

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ
СІКОРСЬКОГО»

ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ



Тема : Оцінка кредитоспроможності фізичних осіб на основі динамічної мережі Байєса

Виконав:

студент 4 курсу
групи КА-43

Макогон Роман Олександрович

Науковий керівник:

к.т.н., старший викладач
каф. ММСА

Терентьєв Олександр Миколайович

КИЇВ - 2018

- **Метою** дослідження є створення програмного продукту для побудови структури дискретної мережі Байеса за навчальними даними, використовуючи евристичний метод.
- **Об'єктом** дослідження є статистичні дані щодо кредитної історії фізичних осіб, які потребують обробки з метою виявлення невідомих взаємозв'язків між змінними.
- **Предметом** дослідження є методи навчання ймовірнісних мереж Байеса за початковими даними та формування ймовірнісного висновку в них.

Актуальність дослідження

- Технологія інтелектуального аналізу даних сьогодні набуває найбільш широкого застосування. Процес обробки сирих даних є суттєвим, бо не інтерпретована інформація не несе нічого корисного.
- За попередні роки були накопичені значні масиви даних стосовно досвіду кредитування фізичних осіб в Україні. Мережі Байєса – ефективний інструмент для роботи з ними.
- Застосування мереж Байєса дозволить підвищити ефективність прийняття рішень щодо видачі кредитів та знизити кількість неповернених позик.

Постановка задачі

Досягнення поставленої мети вимагає розв'язання таких задач:

- Огляд теоретичних основ мереж Байеса
- Аналіз принципів побудови мереж Байеса за навчальними даними.
- Аналіз методів побудови ймовірнісного висновку в мережах Байеса.
- Реалізація евристичного алгоритму побудови мереж Байеса у вигляді програмного продукту.
- Апробація програмного продукту на навчальних даних, що становлять вибірку анкетних даних фізичних осіб при їх кредитуванні.
- Аналіз ефективності створення моделей, порівняння отриманих структур з результатами програм GeNIe та BayesiaLab.

ОСНОВНІ ПОНЯТТЯ

- Мережа Баєса – це пара $\langle G, B \rangle$, де G є спрямованим ациклічним графом, а B є множиною параметрів, які визначають цю мережу. Параметри мережі – множина умовних ймовірнісних розподілів

$$\Theta_{x^{(i)}|pa(x^{(i)})} = P(X^{(i)}|pa(X^{(i)}))$$

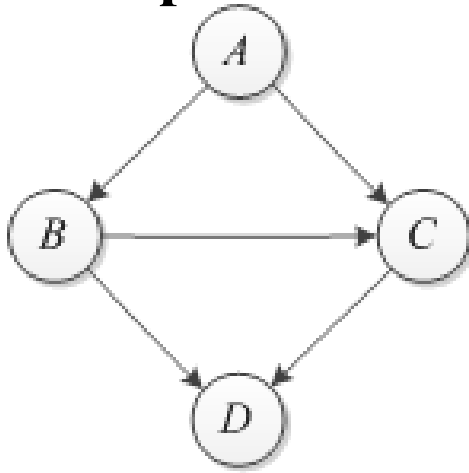
де $Pa(X^{(i)})$ - множина батьків змінної $X^{(i)} \in G$.

- Повна спільна ймовірність мережі Байєса:

$$P_B(X^{(1)}, \dots, X^{(N)}) = \prod_{i=1}^N P_B(X^{(i)}|Pa(X^{(i)}))$$

Типи мереж Байєса

- **Дискретні мережі Байєса** – це такі мережі, в яких змінні вершин дискретні



Батьки	Стани батьків	
A	A=0	A=1
Стан	Ймовірність	
B=0	0,9	0,1
B=1	0,1	0,9

- **Неперервні МБ** – мережі, в яких змінні вузлів – це неперервні величини. Так, події можуть приймати будь-які стани з деякого діапазону. Умовний розподіл випадкової величини

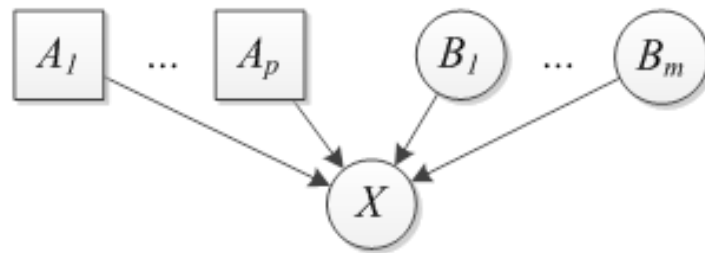
$$f(X | A_i) = N(\mu_x + k_i \cdot \mu_i; \sigma_x) = \frac{1}{\sigma_x \sqrt{2\pi}} \cdot \exp \left[-\frac{1}{2} \left(\frac{x - (\mu_x + k_i \cdot \mu_i)}{\sigma_x} \right)^2 \right]$$

де X має деяку множину батьків $A = \{A_1, \dots, A_n\}$ і стандартне відхилення σ_x

Типи мереж Байєса

Слайд 7 із 35

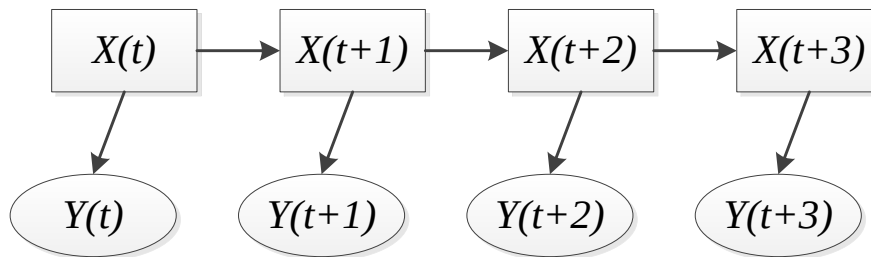
- **Гібридні МБ** – це такі мережі, вершини яких містять і дискретні і неперервні змінні.



У цих мережах Байєса дискретні вершини не мають неперервних батьків, розподіл неперервної змінної X із дискретними батьками A і неперервними батьками B є нормальним:

$$P(X | A = a, B = b) = N(\mu_x(\mu_A, \mu_B); \sigma_x(\sigma_A))$$

- **Динамічні** мережі Байєса є мережами, в яких значення вершин із часом змінюється. Для опису поточного стану процесу можна також додавати приховані вузли. Найпростіший тип динамічної мережі Байєса – це прихована модель Маркова.



Формула Байєса

Правило Байєса для двох подій:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

В термінах гіпотез H та спостережень E :

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E|H) \cdot P(H) + P(E|\neg H) \cdot P(\neg H)}$$

При визначених апіорних імовірностях $P(H)$ та $P(E|H)$ отримуємо апостеріорну імовірність $P(H|E)$.

$$P(H_k|E) = \frac{P(E|H_k) \cdot P(H_k)}{\sum_{i=1}^n P(E|H_i) \cdot P(H_i)}$$



Ймовірнісний висновок

Завданням імовірнісного висновку є пошук і знаходження $P(X | E)$, тобто апостеріорної ймовірності шуканих вершин X при певному значенні спостережуваних вершин E .

Алгоритми точного висновку

1. розповсюдження повідомлення
для однозв'язних мереж (Перла)
2. визначеного перетину
3. виключення змінних
4. кластеризації
5. Інші

Апроксимаційні алгоритми

1. стохастичної вибірки
2. неповного висновку
3. варіаційні
4. пошукові.

Ймовірнісний висновок

У випадку динамічної мережі Байеса, застосовують наступні методи:

Алгоритми точного висновку

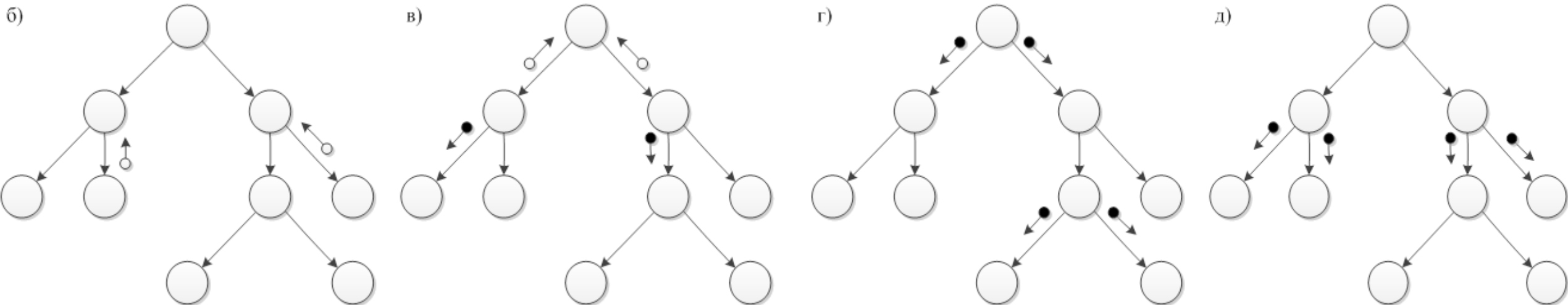
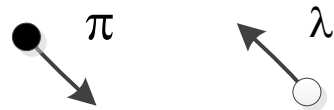
1. Алгоритм прямого і зворотного проходу
2. Граничний алгоритм
3. Інтерфейсний алгоритм
4. Алгоритм із використанням побудови зв'язних дерев

Апроксимаційні алгоритми

1. Детерміністичні
2. Стохастичні

Алгоритм розповсюдження повідомлення

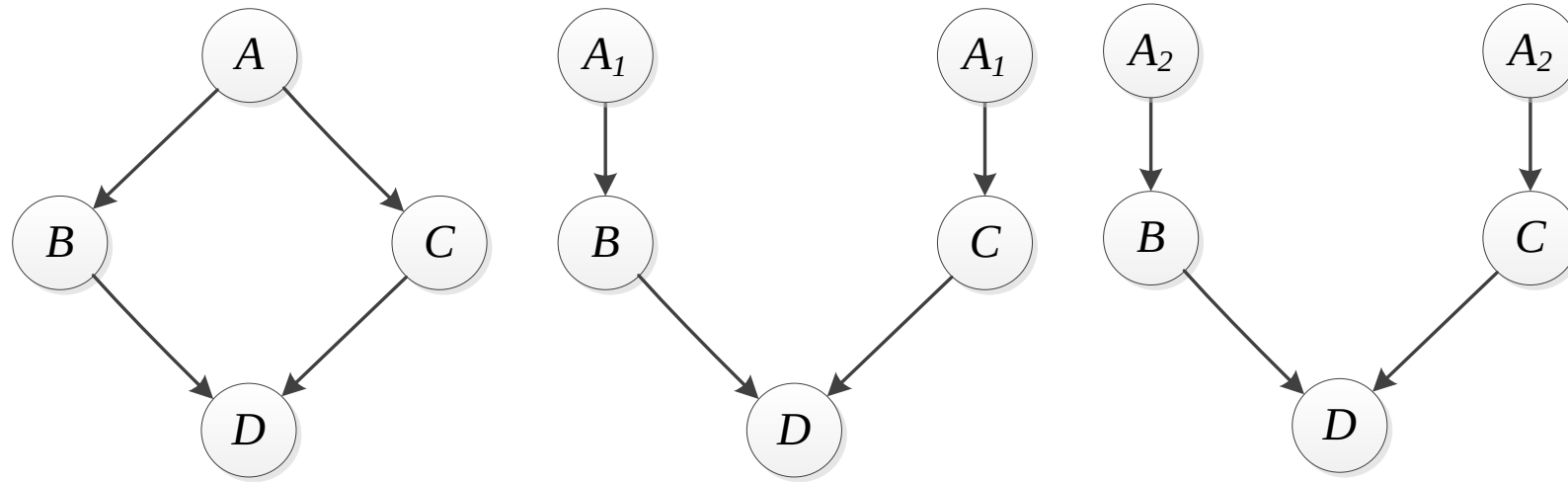
- Вплив нової порції спостережень має розглядатися як збурення. За допомогою пересилання повідомлень між сусідніми вершинами це збурення поширюється по мережі. Інформація щодо стану вершини-батька передається за допомогою π -повідомлення, а щодо стану вершини-нащадка – λ -повідомлення



Алгоритм визначеного перетину

Слайд 12 із 35

- Основна концепція алгоритму – в зміні структури багатозв'язної мережі декількома однозв'язаними за допомогою інстанціювання вершин, входящих до перетину.



Формула, для обчислення сукупного ймовірнісного розподілу мережі може бути записана в такому вигляді:

$$P(A, B, C, D) = P(A = A_1, B, C, D) \cdot P(A = A_1) + P(A = A_2, B, C, D) \cdot P(A = A_2).$$

Ступінь складності алгоритму залежить від кількості інстанціювань, які необхідно обробити й росте експоненційно $O(s \cdot n)$, де s – розмір визначеного перетину, а n – це кількість станів, які можуть приймати вершини перетину.

АЛГОРИТМ ВИКЛЮЧЕННЯ ЗМІННИХ

Слайд 13 із 35

- Основна концепція алгоритму – в обчисленні ймовірності вершини за формулою:

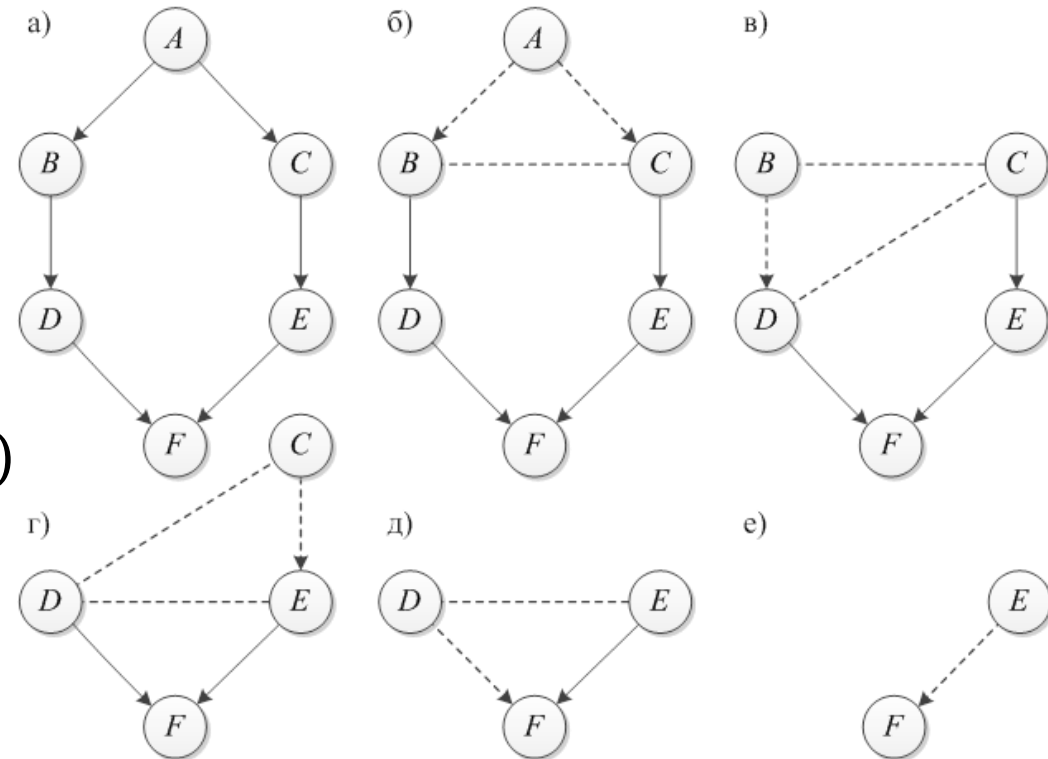
$$P(X_n, e) = \sum_{x_k} \dots \sum_{x_3} \sum_{x_2} \prod_i P(X_i | Pa(X_i))$$

яка базується на формулі декомпозиції

сукупного розподілу ймовірностей мережі:

$$\begin{aligned} P(X_1, \dots, X_n) &= P(X_n | X_1, \dots, X_{n-1}) \\ &\cdot P(X_{n-1} | X_1, \dots, X_{n-2}) \dots P(X_1) = \\ &= \prod_{i=1}^n P(X_i | X_1, \dots, X_{i-1}) = \prod_{i=1}^n P(X_i | Pa(X_i)) \end{aligned}$$

де $Pa(X_i)$ – множина вершин-батьків вершини X_i .



Невідома структура, повні спостереження

- Для побудови структури мережі за навчальними даними використовувався евристичний алгоритм
- Для зменшення складності задачі побудови мережі Байеса формувалась впорядкована множина вершин на основі методу оцінки взаємозв'язку між вершинами – значення взаємної інформації

$$MI(x^i, x^j) = \sum_{x^i, x^j} P(x^i, x^j) \cdot \log\left(\frac{P(x^i, x^j)}{P(x^i)P(x^j)}\right)$$

- Використання функції опису мінімальною довжиною як оціночної функції

Функція опису мінімальною довжиною

Слайд 15 із 35

- Значення функції ОМД довільної вершини в структурі g при заданій послідовності з n спостережень обчислюється за формулою:

$$L(j, g, x^n) = H(j, g, x^n) + \frac{k(j, g)}{2} \cdot \log(n)$$

H означає емпіричну ентропію, k – кількість незалежних умовних ймовірностей

$$H(j, g, x^n) = \sum_{(j, g)} \sum_{q \in A(j)} -n[q, s, j, g] \cdot \log\left(\frac{n[q, s, j, g]}{n[s, j, g]}\right)$$
$$k(j, g) = (\alpha^{(j)} - 1) \prod_{k \in \emptyset(j)} \alpha^k$$

Евристичний алгоритм побудови БМ

Слайд 16 із 35

Вхідні дані: Множина навчальних даних $D = \{d_1, \dots, d_n\}$ $d_i = \{x_i^{(1)} x_i^{(2)} \dots x_i^{(N)}\}$

n – кількість спостережень; N – кількість змінних

Перший етап. Побудова впорядкованої множини вершин

$$V = \{(x^{s1}, x^{s2}); (x^{s3}, x^{s4}); (x^{s5}, x^{s6}); \dots\}$$

Другий етап.

Крок 1. На основі 1-го значення впорядкованої множини вершин будується множина моделей G :

$$\{(s_1 \rightarrow s_2); (s_1 \leftarrow s_2); (s_1 \text{ не залежить від } s_2)\}$$

Крок 2. Пошук серед моделей множини G . Оптимальна структура має найменше значення функції

1. $g^* \leftarrow g_0 (\in G);$

2. для $\forall g \in G - \{g_0\}$ якщо $L(g, x^n) < L(g^*, x^n)$ то $g^* \leftarrow g;$

3. розв'язком є g^* .

Перехід на ітерацію 2.

Умова завершення алгоритму. Перегляд $\frac{n(n-1)}{2}$ елементів впорядкованої множини вершин

- Динамічні мережі Байєса (ДБМ) – розширення статичних мереж Байєса для моделювання розподілів імовірностей на множині випадкових змінних $Z_1, Z_2, \dots$. Як правило, змінні розбиваються на трійки $Z_t = (U_t, X_t, Y_t)$ що позначають множину змінних вхідного, прихованого та вихідного шарів моделі у просторі станів. Розглядається модель стохастичного процесу з дискретним часом, тобто індекс t зростає з появою кожного нового спостереження.
- ДБМ є парою $(B_1, B_{\rightarrow})$, де B_1 – мережа Байєса, що визначає апріорну ймовірність $P(Z_1)$, а $B_{\rightarrow}$ – двошарова мережа Байєса, що визначає $P(Z_t | Z_{t-1})$ за допомогою спрямованого ациклічного графа таким чином:

$$P(Z_t | Z_{t-1}) = \prod_{i=1}^N P(Z_t^i | Pa(Z_t^i))$$

де Z_t^i – i -й вузол в момент часу t , що може бути компонентою X_t, Y_t або U_t , а $Pa(Z_t^i)$ – батьки Z_t^i на графі.

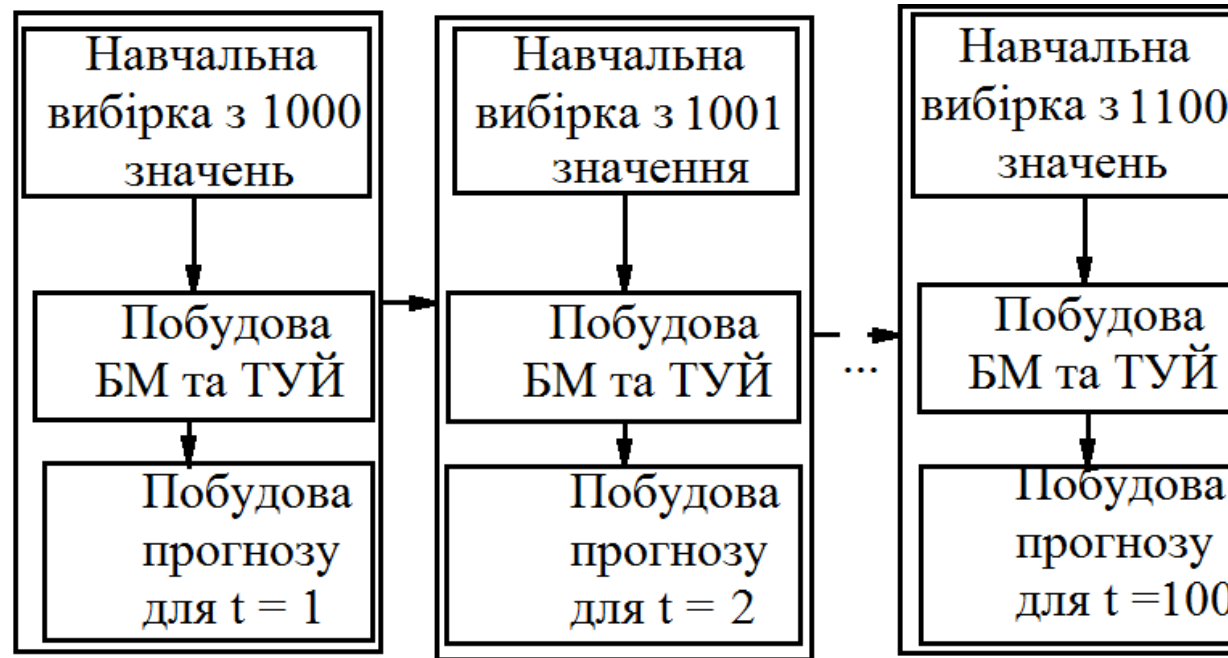
- Дуги між шарами спрямовуються зліва направо, що позначає хід часу.
- Припустимо, що параметри розподілу умовних імовірностей інваріантні в часі, тобто модель є часово-гомогенною. Якщо параметри можуть змінюватися, їх можна додати до простору станів та розглядати, як випадкові змінні.
- Семантика ДБМ передбачає поширення двошарової мережі Байєса на T часових шарів. Результируючий розподіл записується наступним чином:

$$P(Z_{1:T}) = \prod_{t=1}^T \prod_{i=1}^N P(Z_t^i \mid Pa(Z_t^i))$$

- Різниця між ДБМ і прихованою моделлю Маркова полягає в тому, що ДБМ представляє приховані стани як набір випадкових змінних, $X_t^1 \dots X_t^N$, тобто використовує розподілене представлення стану. Натомість в прихованій марковській моделі простір станів складається з однієї випадкової змінної X_t

Динамічні мережі Байєса

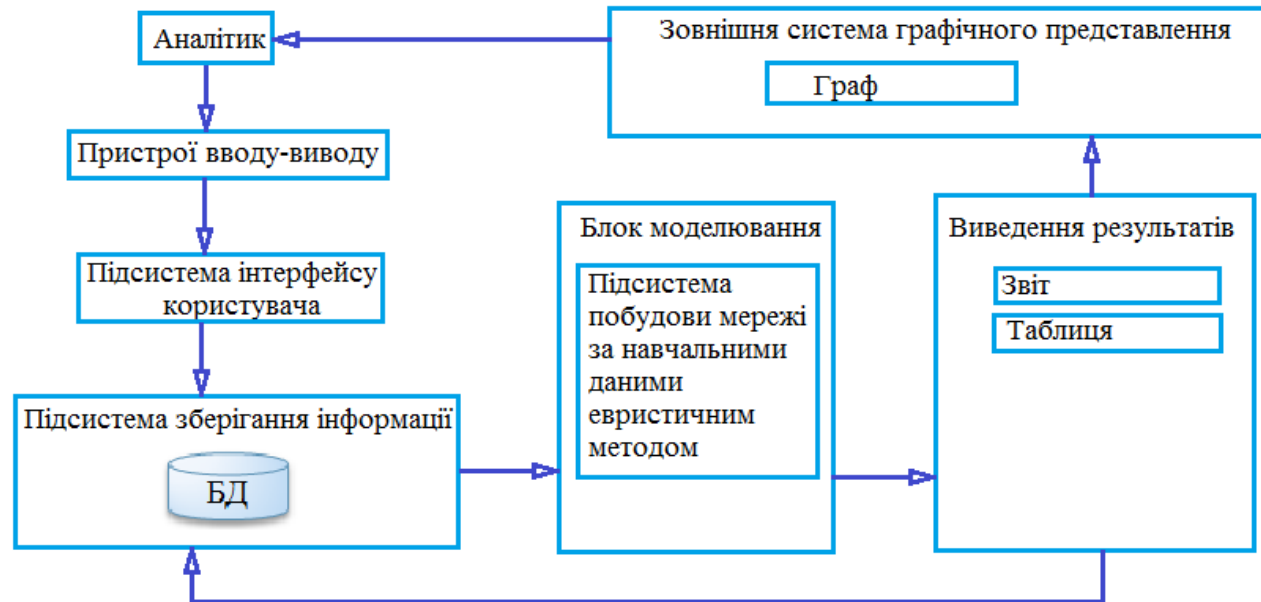
- Для побудови мереж з використанням створеного програмного продукту була використана наступна схема:



Це дозволяє корегувати побудовану модель для врахування надходження даних про нових позичальників банку.

Розроблений програмний продукт

Програмний продукт для побудови структур мереж Байеса за навчальними даними реалізовано мовою програмування SAS Base в середовищі SAS 9.3 TS Level 1M2.



Користувач може завантажити власні дані до програми. Відбувається збереження інформації у базі даних, що накопичує проміжні дані розрахунків, разом з результатами для їх подальшого аналізу та використання. Результати представляють набір таблиць та звітів для прийняття рішення експертом.

Опис набору даних

Для тестування програмного продукту на реальних статистичних даних була використана вибірка даних російського банку ООО «Хоум Кредит энд Финанс Банк», який кредитує фізичних осіб. Набір даних має 597330 записів про клієнтів, з різними типами кредитів та станом їх погашення або заборгованості, та включає в себе інформацію щодо 34 типів анкетних даних по кожній особі. На етапі попередньої обробки даних було відібрано 9 з них.

Назва змінної	Опис
DEF	цільова змінна
CREDIT_ACTIVE	статус кредитного договору
CREDIT_CURRENCY	валюта кредиту
CREDIT_DAY_OVERDUE	поточна прострочена заборгованість, дні
AMT_CREDIT_SUM	сума кредиту
AMT_CREDIT_SUM_DEBT	сума боргу, що залишився
AMT_CREDIT_SUM_OVERDUE	поточна прострочена заборгованість, сума
CREDIT_TYPE	тип договору
AMT_ANNUITY	сума щомісячного платежу

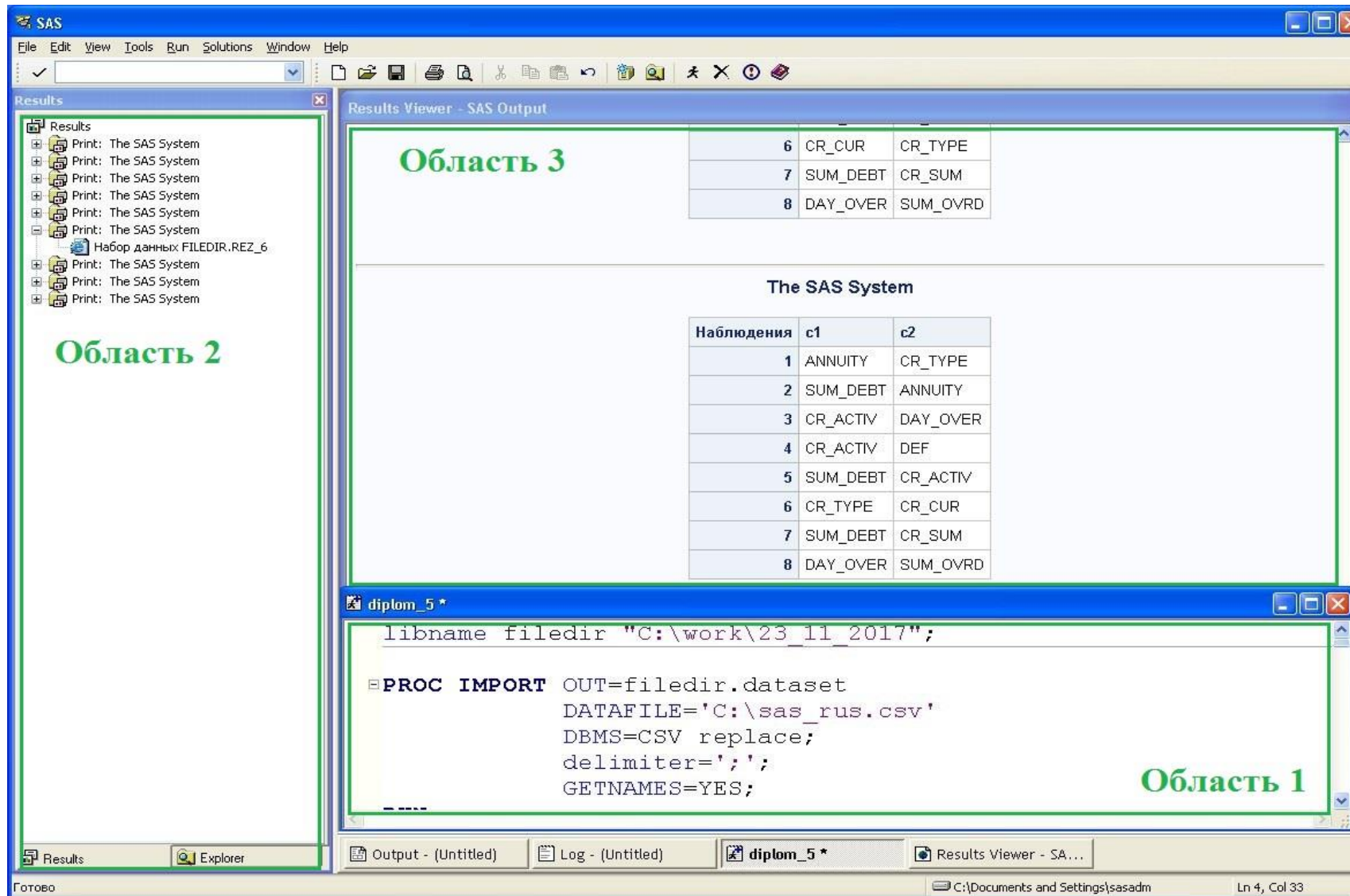
DEF	CREDIT_ACTIVE	CREDIT_CURRENCY	CREDIT_DAY_OVERDUE	AMT_CREDIT_SUM	AMT_CREDIT_SUM_DEBT	AMT_CREDIT_SUM_OVERDUE	CREDIT_TYPE	AMT_ANNUITY
0	1	rur	0	27890	4768	0	5	0
0	1	rur	0	34860	28079	0	5	2799
0	0	rur	0	4090	0	0	5	0
0	1	rur	0	15342	7587	0	5	1448
0	0	rur	0	8495	0	0	5	653
0	2	rur	2601	8824	0	0	5	1103

Також було проведено дискретизацію початкових даних

DEF	CREDIT_ACTIVE	CREDIT_CURRENCY	CREDIT_DAY_OVERDUE	AMT_CREDIT_SUM	AMT_CREDIT_SUM_DEBT	AMT_CREDIT_SUM_OVERDUE	CREDIT_TYPE	AMT_ANNUITY
0	1	rur	<=19.015	<=227169	<=86893	<=3264.31	5	<=4862
0	1	rur	<=19.015	<=227169	<=86893	<=3264.31	5	<=4862
0	0	rur	<=19.015	<=227169	<=86893	<=3264.31	5	<=4862
0	1	rur	<=19.015	<=227169	<=86893	<=3264.31	5	<=4862
0	0	rur	<=19.015	<=227169	<=86893	<=3264.31	5	<=4862
0	2	rur	<=2871.265	<=227169	<=86893	<=3264.31	5	<=4862

Головне вікно програми

Слайд 23 із 35

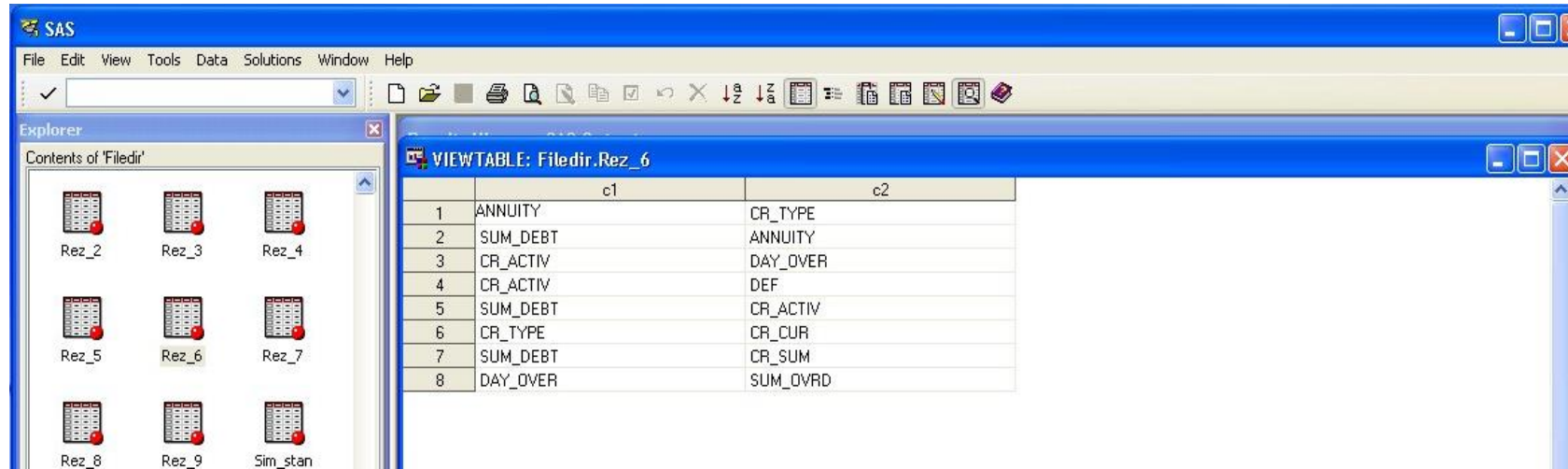


(«Область 1»),
користувач може зазначити
шлях до розташування
вхідних даних, тип файлу,
шлях збереження
результатів.

(«Область 2»),
користувач може аналізувати
збережені під час роботи
файли.

(«Область 3»),
користувач може
переглянути звіт з
результатами.

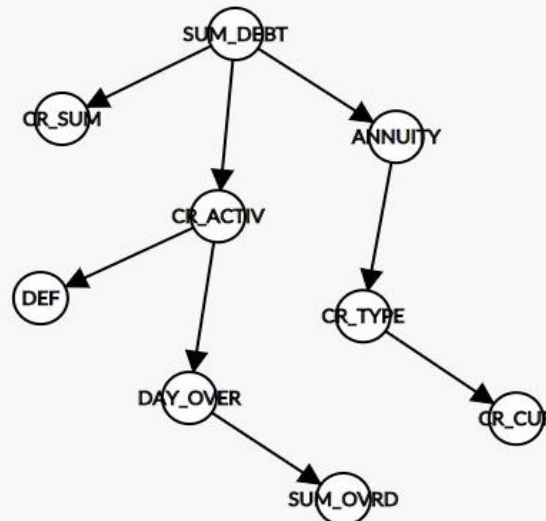
Результати роботи



The screenshot shows the SAS interface. On the left, the 'Explorer' window displays a file tree with folders 'Rez_2' through 'Rez_9' and a file 'Sim_stan'. The main window, titled 'VIEWTABLE: Filedir.Rez_6', displays a table with two columns, 'c1' and 'c2', containing 8 rows of data.

	c1	c2
1	ANNUITY	CR_TYPE
2	SUM_DEBT	ANNUITY
3	CR_ACTIV	DAY_OVER
4	CR_ACTIV	DEF
5	SUM_DEBT	CR_ACTIV
6	CR_TYPE	CR_CUR
7	SUM_DEBT	CR_SUM
8	DAY_OVER	SUM_OVRD

Node Count:
9
Graph Data:
ANNUITY CR_TYPE
SUM_DEBT ANNUITY
CR_ACTIV DAY_OVER
CR_ACTIV DEF
SUM_DEBT CR_ACTIV
CR_TYPE CR_CUR
SUM_DEBT CR_SUM
DAY_OVER SUM_OVRD



Слайд 24 із 35

Результатом роботи програми є матриці суміжності графу, які можна візуалізувати з допомогою ресурсу CSAcademy Graph Editor. На тестовому датасеті було побудовано 9 структур, що мають мінімальне значення ОМД, тобто є оптимальними з точки зору реалізованого методу побудови мереж Байеса. Мережі, отримані в результаті, попарно відрізняються один від одного напрямками ребер між окремими вершинами.

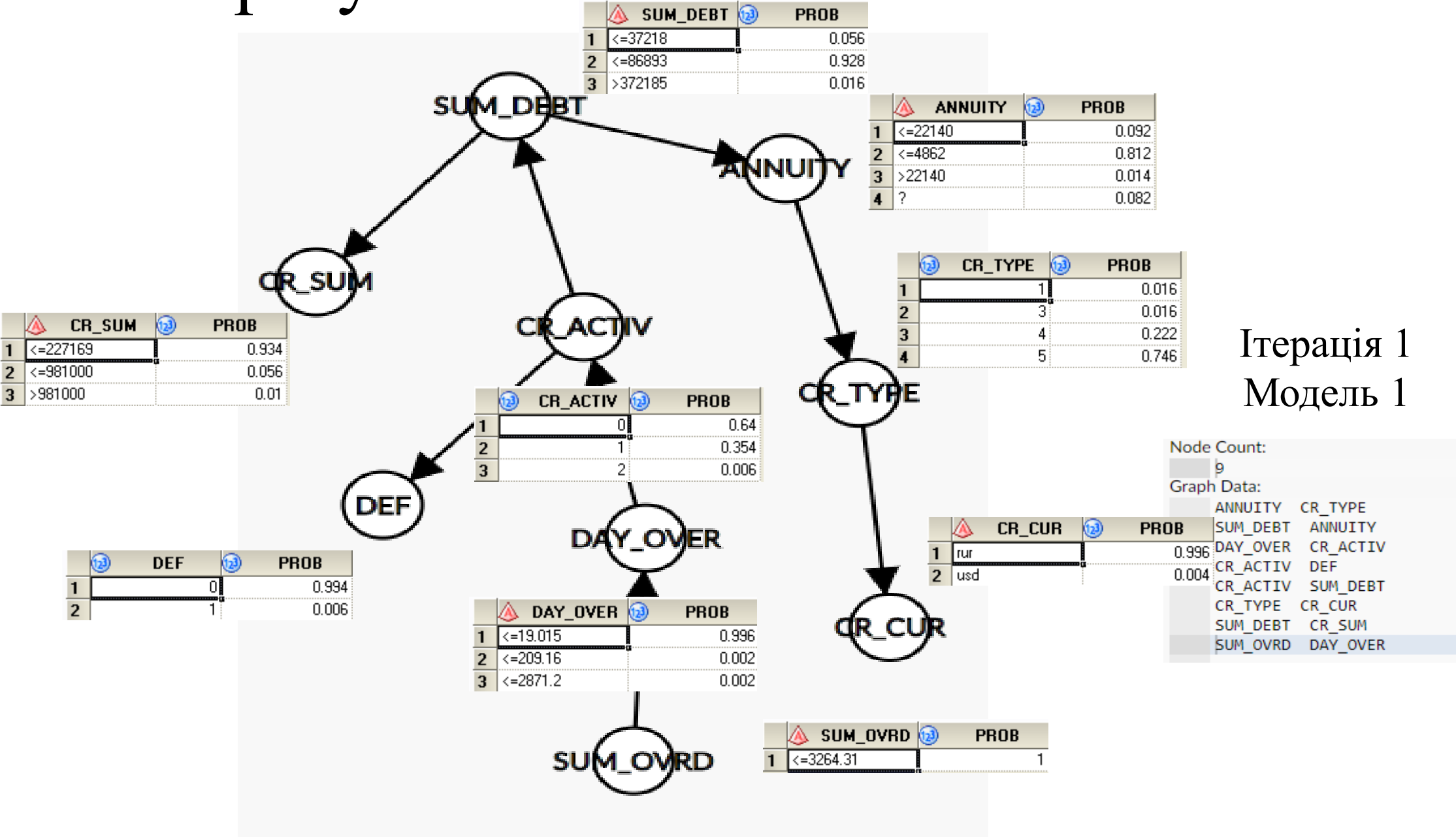
Аналіз результатів

Слайд 25 із 35

Для порівняння якості побудованої моделі в програмному продукті з використанням евристичного методу, було побудовані моделі в системі GeNIe 2.2.2601.2 (64-bit) та BayesiaLab 7.0.1 DE. Ці дві програми пропонують функціонал для побудови мереж Байеса за навчальними даними.

Створений програмний продукт побудував 9 оптимальних моделей, з яких модель номер 5 повністю співпала з результатом, запропонованим програмою GeNIe, модель номер 6 була ідентична отриманій в програмі BayesiaLab.

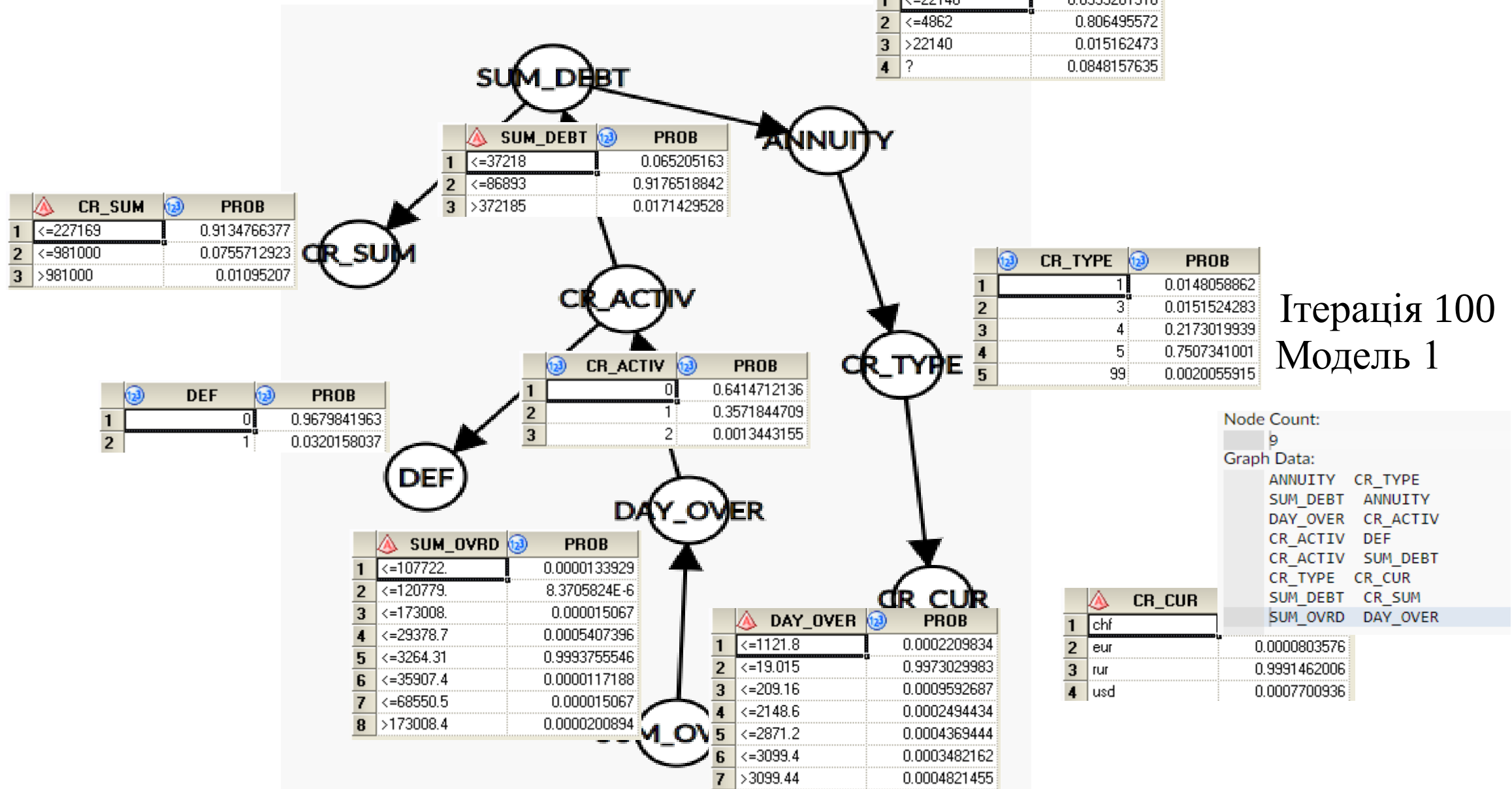
Аналіз результатів



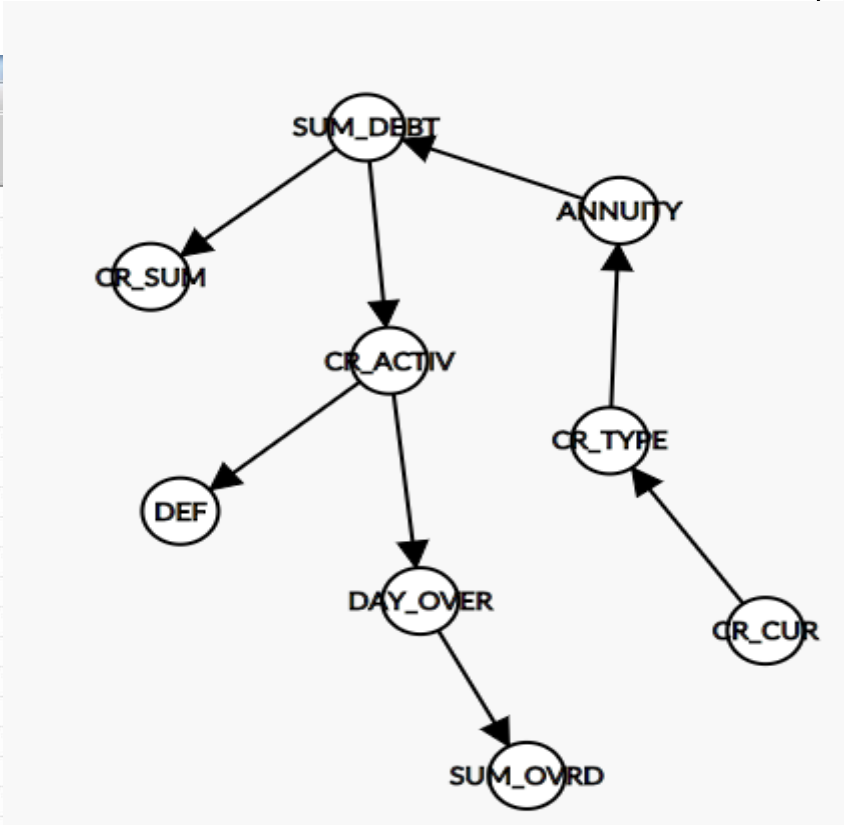
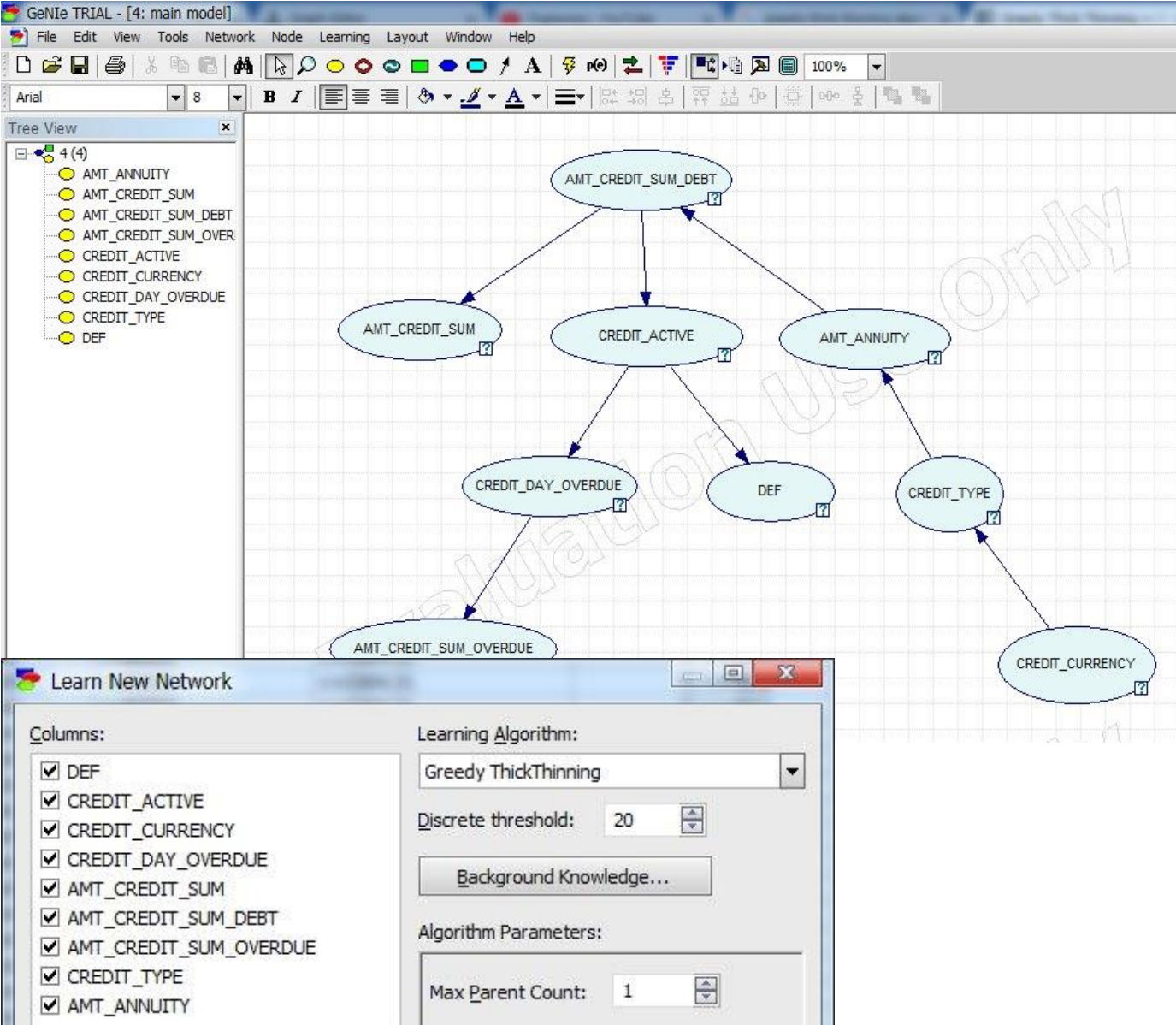
Ітерація 1
Модель 1

Аналіз результатів

Слайд 27 із 35



Аналіз результатів



Модель 5
Ітерація 100

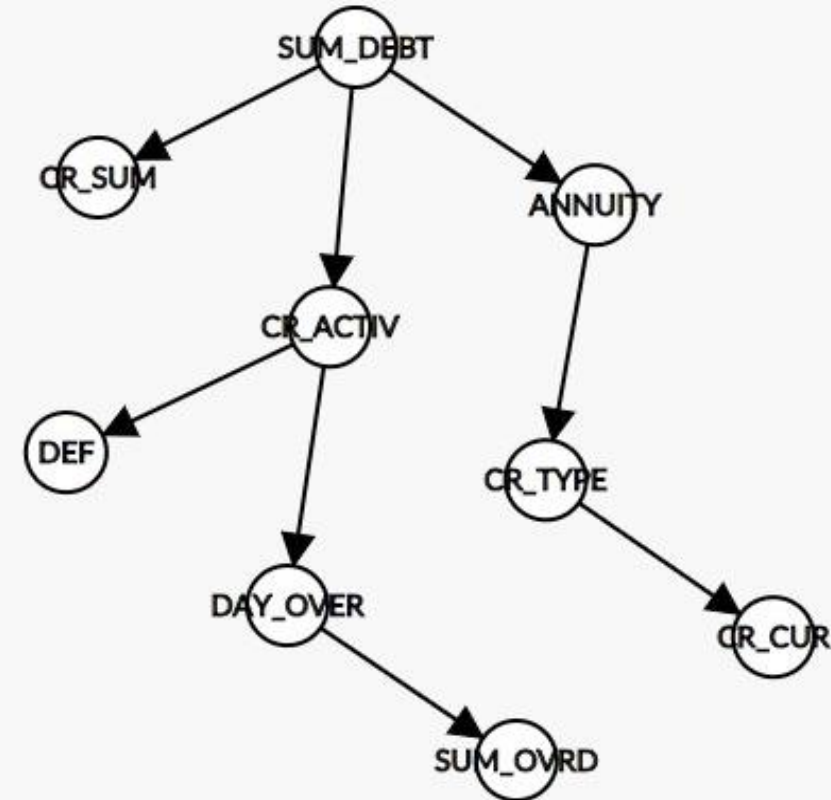
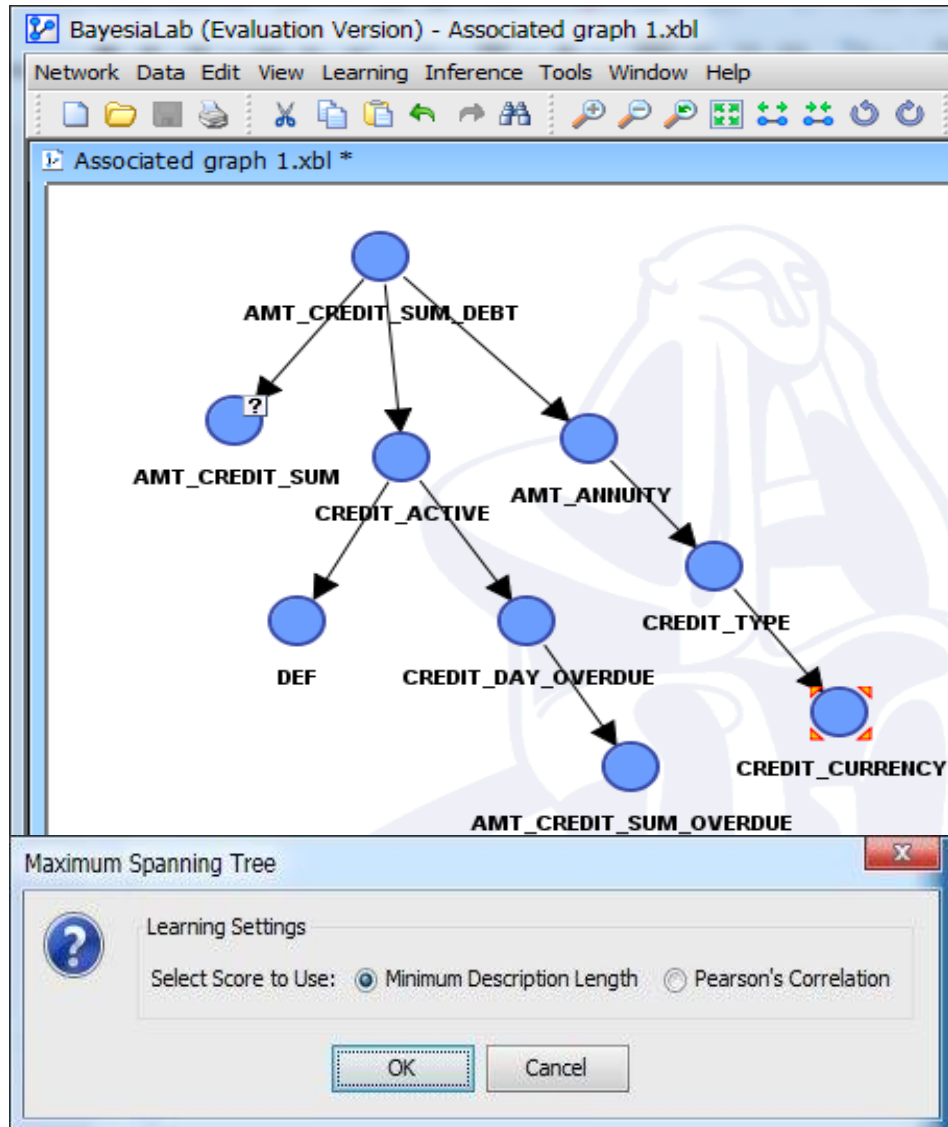
Node Count: 9

Graph Data:

CR_TYPE	ANNUTY
ANNUTY	SUM_DEBT
CR_ACTIV	DAY_OVER
CR_ACTIV	DEF
SUM_DEBT	CR_ACTIV
CR_CUR	CR_TYPE
SUM_DEBT	CR_SUM
DAY_OVER	SUM_OVRD

Аналіз результатів

Слайд 29 із 35



Модель 6
Ітерація 100

Node Count:
9

Graph Data:

ANNUITY	CR_TYPE
SUM_DEBT	ANNUITY
CR_ACTIV	DAY_OVER
CR_ACTIV	DEF
SUM_DEBT	CR_ACTIV
CR_TYPE	CR_CUR
SUM_DEBT	CR_SUM
DAY_OVER	SUM_OVRD

Оцінювання якості побудови мереж Байєса

Для обчислення структурної різниці двох мереж A і B використовують формулу симетричної різниці структур:

$$\begin{aligned}\delta &= \sum_{i=1}^n \delta_i = \sum_{i=1}^n \text{card}\left(\Pi^{(i)}(B) \Delta \Pi^{(i)}(A)\right) = \\ &= \sum_{i=1}^n \text{card}\left(\left(\Pi^{(i)}(B) \setminus \Pi^{(i)}(A)\right) \cup \left(\Pi^{(i)}(A) \setminus \Pi^{(i)}(B)\right)\right)\end{aligned}$$

$\Pi^{(i)}(A)$ – множина предків вершини i мережі A .

Необхідною умовою застосування критерію є однакова кількість вершин n мереж A і B .

Аналіз результатів

У таблиці наведені показники структурної різниці, оцінені за кількістю доданих, видалених зв'язків та зв'язків із зміненим напрямком. Результати підтвердили ефективність побудови моделей у порівнянні зі світовими аналогами.

Мережа 1	Мережа 2	Структурна Різниця
Побудована мережа 1	GeNIe	12
Побудована мережа 5	GeNIe	0
Побудована мережа 6	BayesiaLab	0
GeNIe	BayesiaLab	6
Побудована мережа 1	BayesiaLab	6

Висновки

Було розглянуто базові поняття мереж Байєса, теорії графів. Також були розглянуті основні типи мереж, принципи побудови МБ за навчальними даними. Розглянута задача побудови ймовірнісного висновку, алгоритми побудови такого висновку.

Було спроектовано систему підтримки прийняття рішень з використанням евристичного алгоритму, для побудови мереж Байєса за навчальними даними.

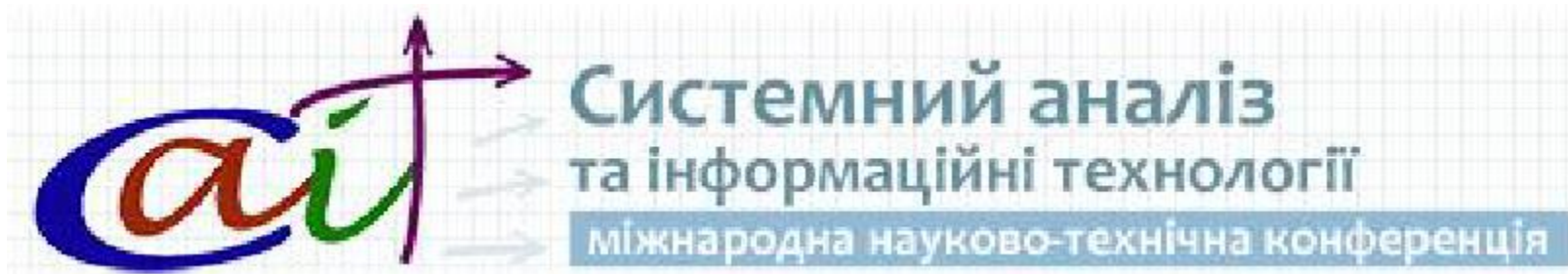
Продукт адаптований для широкого кола початкових наборів даних, їх типів, внутрішньої структури та розміру. Розроблений програмний продукт апробовано на реальних статистичних даних з банку ООО «Хоум Кредит энд Финанс Банк». Було проведено порівняльний аналіз результатів методу з результатами роботи програм GeNIe 2.2.2601.2 (64-bit) та BayesiaLab 7.0.1 DE на цьому ж датасеті. Результати підтвердили ефективність побудови моделей.

Перспективи щодо подальших досліджень

- розробка методу знаходження імовірнісного висновку у побудованих моделях.
- розробка власного блоку дискретизації даних

Публікації

- участь у XX Міжнародній науково-технічній конференції «Системний аналіз та інформаційні технології» - 2018 рік.



- участь у Всеукраїнському конкурсі наукових робіт SAS Україна – 2017 рік.

(<https://sites.google.com/site/sas4ua2017/home>)



Дякую за увагу !